Lexicons Gain the Upper Hand in Arabic MWE Identification

Najet Hadj Mohamed[12], Agata Savary[3], Cherifa Ben Khelil[1] , Jean-Yves Antoine[1], Iskander Keskes[2], Lamia Belguith Hadrich[2]

1 University of Tours, LIFAT, ICVL; 2 University of Sfax MIRACL; 3 University of Paris-Saclay, LISN

MWE identification

- **MWE in MSA :**
  - Tow words with unexpected behavior [1]

ملحه علي ركبته [His salt on his knees] ≠ Salt + Knees

- **Objectif:**

Identifying MWEs using an Arabic lexicon (capturing unseen expressions more effectively and reducing the ambiguity of literal interpretations)

- **Challenges:**
  - **Unseen VMWEs:** Identifying MWEs that have not been previously encountered in training datasets.
  - **Idiomatic Ambiguity:** Differentiating literal from figurative meanings.

APPROACH

STEP 1

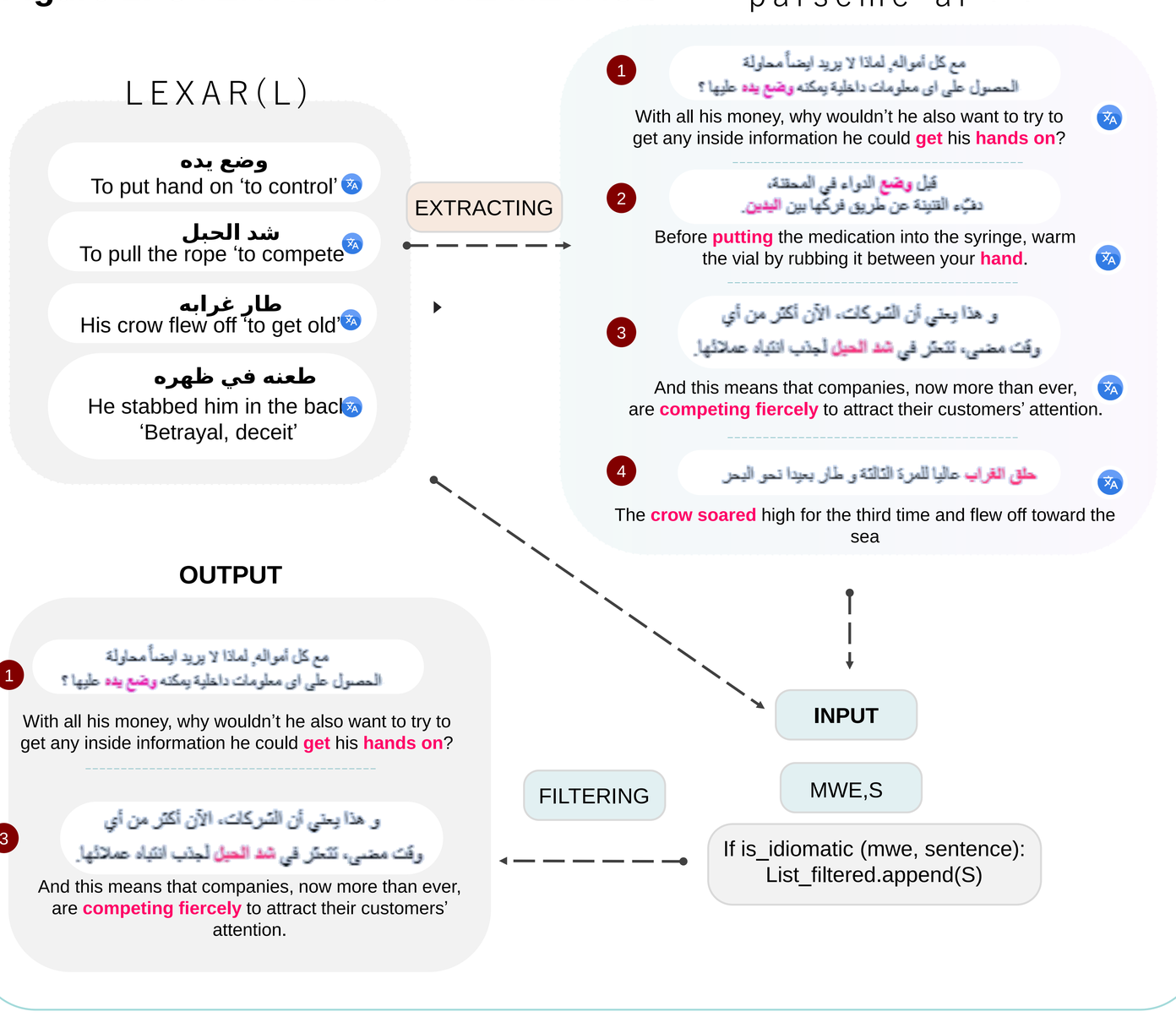
Identifying VMWE candidates
(based on lemmas associated with each MWE lexicon)

STEP 2

Disambiguating candidate VMWE occurrences:
PIEC: Potential Identiomatic Expression Classifier

Architecture:
1. **Tokenization:** The input sequence SS and the target PIE are tokenized.
2. **Embedding Generation:** BERT is used to generate contextual embeddings, providing vector representations for both the PIE and its context SS.
3. **Feature Extraction:** A Bidirectional LSTM (BiLSTM) layer extracts features from these embeddings, resulting in $h(S)$=BiLSTM($e(S)$)$h(S)$=BiLSTM($e(S)$) and $h(PIE)$=BiLSTM($e(PIE)$)$h(PIE)$=BiLSTM($e(PIE)$).
4. **Attention Flow:** The attention flow layer integrates context and query information, producing query-aware vector representations and fusing $h(S)h(S)$ and $h(PIE)h(PIE)$ into a cohesive contextual representation.
5. **MaxPooling:** A MaxPooling layer reduces the dimensions of the data while retaining key features.
6. **Classification:** The integrated representation is passed through Dense layers, with the final classification performed using a sigmoid layer.

Figure 1: OVERVIEW OF THE METHOD



EVALUATION

DATASET

MWE identification data

corpus from PARSEME 1.3 [1-2]

parseme-ar:
4,7000 VMWEs within 7,500 sentences.

VMWE lexicon
1504 Arabic VMWE manuually annotated.
From "Contextual Dictionary of Idiomatic Expressions" by [5]

Idiomatic expression classifier

MAGPIE corpus [3]
English

COLF-VIDcorpus [4]
German

Generated by ChatGPT
Arabic

Lang	Literal	Figurative	Total
AR-train	103	202	305
AR-dev	16	30	46
AR-test	29	57	86
COLF-VID-train	1,172	5,705	6,902
COLF-VID-dev	264	1,214	1,488
COLF-VID-test	265	1,238	1,511
MAGPIE-train	2,676	12,676	15,352
MAGPIE-dev	595	2719	3314
MAGPIE-test	635	3339	3974

Table1 :Literal and idiomatic occurrences of PIEs in Arabic (AR), German (DE) (we excluded both the types of BOTH and UNDECIDABLE, which accounts for the disparity in the count between literal and idiomatic expressions compared to the total) and English(EN)

RESULTS

Figure 2: Comparing our approach performance with MTLB-STRUCT on MWE-based and unseen MWE-based metrics.

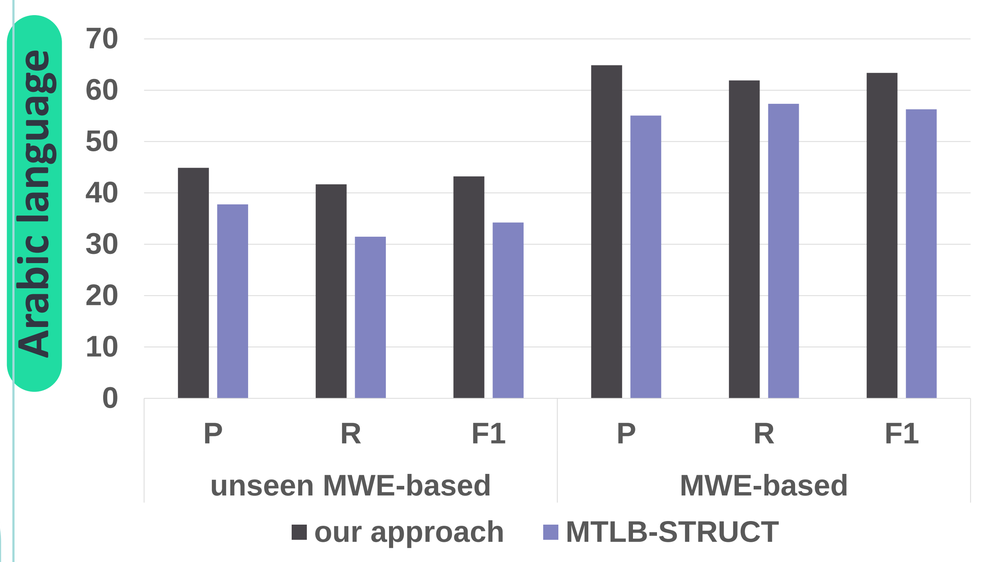
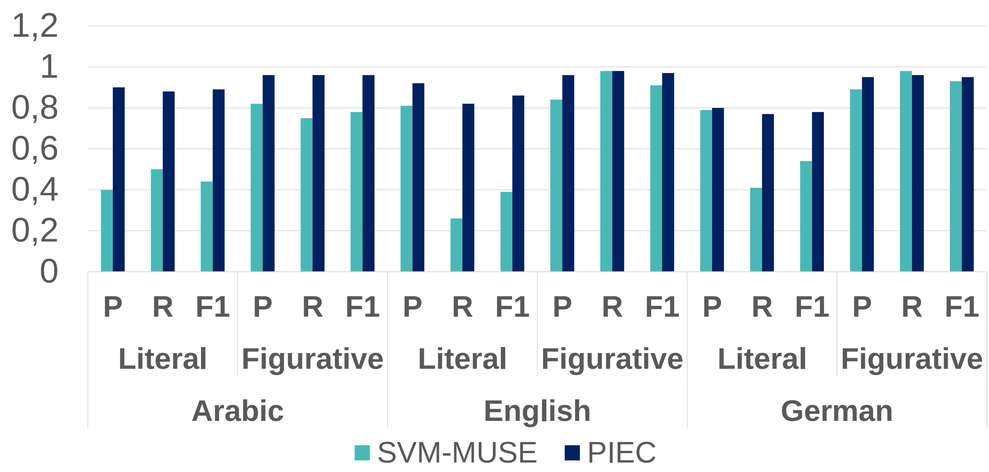


Figure 3: Comparing our approach performance with MTLB-STRUCT on MWE-based and unseen MWE-based metrics.



Discussion:

Identification of VMWE candidates:

- Our approach outperforms MTLB-STRUCT outperforms MTLB-STRUCT in terms of MWE-based F1 score by 7% and for unseen MWEs by 9% (see Figure 2)
- Among the 278 unseen VMWEs assessed, our approach detected 125, whereas MTLB-STRUCT identified 104 out of the total.

Disambiguation (see Figure 3) **:**

- It performs highly better on both literal and figurative class across all languages, even when dealing with unbalanced data in German and English.

REFERENCES

[1] Mohamed, Najet Hadj, et al. "Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure." *13th Conference on Language Resources and Evaluation (LREC 2022)*. 2022.
[2] Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, et al. 2023. Parseme corpus release 1.3. In Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 24–35.
[3] Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 279–287.
[4] Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of german verbal idioms with a bilstm architecture. In Proceedings of the Second Workshop on Figurative Language Processing.
[5] Mahmoud Ismail Elsini. 1998. *Contextual dictionary of idiomatic expressions*. Lebanon Library

ACKNOWLEDGEMENT

Special thanks to Rafael Ehren and Laura Kallmeyer for their support and provision of preprocessed English data. Funding was provided by the UniDive COST Action (CA21167).